Expert Systems With Applications 43 (2016) 82–92

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

An improved global feature selection scheme for text classification

Alper Kursat Uysal*

Department of Computer Engineering, Anadolu University, Eskisehir, Turkiye

ARTICLE INFO

Keywords:
Global feature selection
Filter
Text classification
Pattern recognition

ABSTRACT

Feature selection is known as a good solution to the high dimensionality of the feature space and mostly preferred feature selection methods for text classification are filter-based ones. In a common filter-based feature selection scheme, unique scores are assigned to features depending on their discriminative power and these features are sorted in descending order according to the scores. Then, the last step is to add top-*N* features to the feature set where *N* is generally an empirically determined number. In this paper, an improved global feature selection scheme (IGFSS) where the last step in a common feature selection scheme is modified in order to obtain a more representative feature set is proposed. Although feature set constructed by a common feature selection scheme successfully represents some of the classes, a number of classes may not be even represented. Consequently, IGFSS aims to improve the classification performance of global feature selection methods by creating a feature set representing all classes almost equally. For this purpose, a local feature selection method is used in IGFSS to label features according to their discriminative power on classes and these labels are used while producing the feature sets. Experimental results on well-known benchmark datasets with various classifiers indicate that IGFSS improves the performance of classification in terms of two widely-known metrics namely Micro-F1 and Macro-F1.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Rapid developments of internet technologies lead an increase on the amount of electronic documents worldwide. Consequently, hierarchical organization of these documents becomes a necessity. This situation enhances the importance of text classification whose goal is to classify texts into appropriate classes according to their contents. Text classification is applied to numerous domains such as topic detection (Rill, Reinel, Scheidt, & Zicari, 2014), spam e-mail filtering (Gunal, Ergin, Gulmezoglu, & Gerek, 2006; Idris & Selamat, 2014), SMS spam filtering (Uysal, Gunal, Ergin, & Gunal, 2013), author identification (Zhang, Wu, Niu, & Ding, 2014), web page classification (Saraç & Özel, 2014), and sentiment analysis (Medhat, Hassan, & Korashy, 2014). Text classification tasks can be realized with schemes having different settings. A fundamental text classification scheme, as in many different pattern recognition problems, consists of feature extraction and classification stages. Due to the nature of the problem, feature extraction mechanism needs to extract numerical information from raw text documents. Then, any classifier can be used to finalize the text classification process by predicting the label of documents. However, preprocessing (Uysal & Gunal, 2014) and feature selection (Uysal et al., 2013) are known as very important stages besides feature extraction and classification. Researchers in this field are still studying on enhancing the performance of text classification by incorporating various preprocessing (Dara, Dowling, Travers, Cooper, & Chapman, 2008; Uysal & Gunal, 2014), feature extraction (Vicient, Sánchez, & Moreno, 2013), feature selection (Uysal & Gunal, 2012; Wang, Liu, Feng, & Zhu, 2015), and classification (B. Yang, Zhang, & Li, 2011) methods.

Although there exist some recent studies about improving the feature extraction with the contribution of Wikipedia or similar resources, bag-of-words approach (Joachims, 1997) is the commonly used technique for feature extraction stage. In this approach, the orders of terms are neglected and text documents are represented with weighted frequencies (i.e., TF-IDF (Manning, Raghavan, & Schutze, 2008)) of the unique terms in the collection. As each unique term is used in the construction of the feature set, even a collection including small number of documents may be expressed with thousands of features. Excessive numbers of features may have negative effects on both classification accuracy and computational time. Therefore, most of the researchers concern with the feature selection stage in order to overcome these kinds of negative effects.

Feature selection techniques are generally categorized as filters, wrappers, and embedded methods. While wrappers and embedded methods require a frequent classifier interaction in their flow, filters do not need any classifier interaction during the construction of the feature set. Requirement of a classifier interaction may increase running time and make the feature selection method adapted

* Tel.: +90 2223213550.
E-mail address: akuysal@anadolu.edu.tr

http://dx.doi.org/10.1016/j.eswa.2015.08.050
0957-4174/© 2015 Elsevier Ltd. All rights reserved.

to a specific learning model. Due to these reasons, filter-based methods are preferred more compared to wrappers and embedded methods.

Filter-based methods can be divided into two categories referred as global and local depending on whether they assign a unique score or multiple class-based scores for any feature (Taşcı & Güngör, 2013). In the case of local feature selection methods, a globalization policy is necessary to convert the multiple local scores into a unique global score (Uysal & Gunal, 2012). On the other hand, in the case of global feature selection methods, the scores can be directly used for feature ranking. The features are ranked in descending order and top-*N* features are included in the feature set (Guyon & Elisseeff, 2003) where *N* is usually an empirically determined number. Some examples to global feature selection methods for text classification are document frequency (Yang & Pedersen, 1997), information gain (Lee & Lee, 2006), improved Gini index (Shang et al., 2007), and distinguishing feature selector (Uysal & Gunal, 2012). Another categorization about characteristics of filter-based feature selection methods is whether they are one-sided or two-sided (Ogura, Amano, & Kondo, 2011). In one-sided metrics, while features indicating membership to classes have a score greater than or equal to 0, features indicating non-membership to classes have a score smaller than 0. As features are ranked in descending order and the features having highest scores are included in the feature set, the negative features are not used in case there is no candidate positive feature. However, scores of two-sided methods are greater than or equal to 0. They implicitly combine positive and negative features which indicate the membership and non-membership to any class, respectively. In this case, considering one-against-all strategy in feature selection, positive features attain higher scores than negative ones. Thus, the negative features are rarely added to the feature set in two-sided metrics. Some examples to one-sided feature selection metrics for text classification are odds ratio (Zheng, Wu, & Srihari, 2004) and correlation coefficient (Ogura et al., 2011). In addition to the proposal of new metrics, feature selection studies for text classification proceed with improvement of current feature selection methods and developing ensemble approaches which combine various methods.

In the literature, there exist some studies dealing with integration of negative features in the feature set especially to handle the problems resulting from class imbalances. In previous studies, a local feature selection method which explicitly combines positive and negative features is proposed (Zheng & Srihari, 2003; Zheng et al., 2004). Experimental results on a single dataset show the efficiency of the proposed approach on imbalanced datasets. In a more recent study, the ability of selecting suitable negative features for some local feature selection methods is investigated on imbalanced datasets (Ogura, Amano, & Kondo, 2010). In another study, one-sided and two-sided feature selection metrics are compared for imbalanced text classification (Ogura et al., 2011). In one of the previous studies, a feature selection technique that automatically detects appropriate number of features containing both positive and negative features is proposed (Pietramala, Policicchio, & Rullo, 2012). The performance of the proposed approach which selects dynamic amount of features is compared with the performance of feature sets with some predetermined feature dimensions. The experiments show that the proposed approach succeeds in most of the experiments. Also, a comparison is carried out on two-sided feature selection metrics for text classification and an adaptive feature selection framework is proposed (Taşcı & Güngör, 2013). It is concluded that selecting different number of features for each class improves the performance of classification on imbalanced datasets. Apart from these, there exist some previous text classification studies dealing with combining the power of various feature selection methods. In a study, information gain method is separately combined with genetic algorithm and principal component analysis (Uguz, 2011), respectively. It is reported that both of these combination methods attains better performance than the individual performance of information gain. In a more recent study, several filter methods are combined with genetic algorithm (Gunal, 2012). The results indicate that this combination outperform the individual performances of the filter methods. In this study, contribution ratio of various feature selection metrics into the final feature set is also investigated. Besides, there exist some recent studies proposing solutions to determination of ideal number of features used for representation of documents automatically. As an example, a method that attempts to represent each document in the training set with at least one feature is proposed (Pinheiro, Cavalcanti, Correa, & Ren, 2012). It is stated that this approach obtains equivalent or better results than classical filter-based feature selection methods that attempts to determine the ideal number of features in a trial and error methodology. As another example to this kind of approaches, in a more recent study, representation of documents with more than one feature is proposed in order to improve the performance of classification (Pinheiro, Cavalcanti, & Ren, 2015). It is concluded that this approach performs better than or equal to the former one that each document is represented with only one feature. In addition, an improved feature selection scheme aiming to improve filter-based feature selection methods is proposed (J. Yang, Qu, & Liu, 2014). The main idea behind this study is to consider the imbalance factor of the training sets in the globalization process of class-based feature selection scores. It is reported that this improved scheme can significantly improve the performance of feature selection methods.

In spite of numerous approaches in the literature, feature selection for text classification is still an ongoing research topic. In this study, being inspired from some of the abovementioned studies, a new method namely improved global feature selection scheme (IGFSS), is proposed. IGFSS is a new approach which has some similarities with the characteristics of other approaches in the literature. These similarities can be listed as being a hybrid approach combining the power of two feature selection methods, benefiting from the power of negative features, and proposing a generic solution for all of the filter-based global feature selection methods. IGFSS aims to improve the classification performance of global feature selection methods by creating a feature set representing all classes nearly equally. For this purpose, a one-sided local feature selection method is integrated to the feature selection process besides a global feature selection method. Initially, the one-sided local feature selection method assigns a class label to each feature with a positive or negative membership degree. So, positive and negative features mentioned in the previous works are used as a part of the new method. Odds ratio was employed as one-sided local feature selection method during experiments. Instead of adding top-*N* features having highest global feature selection scores to the feature set, equal number of features representing each class equally with a certain membership and non-membership degree were included in the final feature set. In the experiments, an empirically determined negative feature ratio was used to represent each class with nearly same number of negative features. The experiments were carried out for different classification algorithms, datasets, and success measures. So, effectiveness of IGFSS was observed under different conditions. Results of the experimental analysis revealed that IGFSS offers better performance than the individual performance of global feature selection methods for all cases. In order to analyze classification performances, two common metrics for text classification was employed in the experiments.

Rest of the paper is organized as follows: feature selection methods used in this study are briefly described in Section 2. Section 3 introduces the details of IGFSS method. In Section 4, the classifiers used in the experiments are explained in details. Section 5 presents the experimental study and results which are related to accuracy, for each dataset, classifier, and success measure. Finally, some concluding remarks are given in Section 6.

2. Feature selection methods

Global feature selection methods and one-sided local feature selection methods are within the scope of this paper. As it was pointed out in the previous section, widely-known global feature selection methods are document frequency (Yang & Pedersen, 1997), information gain (Lee & Lee, 2006), Gini index (Shang et al., 2007), and distinguishing feature selector (Uysal & Gunal, 2012). Document frequency is not a part of this study because it does not seem to be successful compared to the other methods. Odds ratio (Forman, 2003) and correlation coefficient (Zheng & Srihari, 2003) can be listed in the category of one-sided local feature selection methods. In this study, odds ratio was utilized as it produces excessive number of negative features (Forman, 2003). Therefore, efficacy of the proposed IGFSS method was assessed on information gain, Gini index and distinguishing feature selector. Mathematical backgrounds of the existing feature selection methods used in this study are provided in the following subsections.

2.1. Information gain (IG)

IG scores show the contribution ratio of the presence or absence of a term to correct classification of text documents (Forman, 2003). IG assigns a maximum value to a term if it is a good indicator for assigning the document to any class. As it is indicated below, IG is a global feature selection metric as producing only one score for any term t and this score is calculated using

$$IG(t) = -\sum_{i=1}^{M} P(C_i)\log P(C_i) + P(t)\sum_{i=1}^{M} P(C_i|t)\log P(C_i|t) + P(\bar{t})\sum_{i=1}^{M} P(C_i|\bar{t})\log P(C_i|\bar{t}), \tag{1}$$

where M is the number of classes, $P(C_i)$ is the probability of class C_i, $P(t)$ and $P(\bar{t})$ are the probabilities of presence and absence of term t, $P(C_i|t)$ and $P(C_i|\bar{t})$ are the conditional probabilities of class C_i given presence and absence of term t, respectively.

2.2. Gini index (GI)

GI is a global feature selection method for text classification which can be defined as an improved version of an attribute selection algorithm used in decision tree construction (Shang et al., 2007). It has a simple formulation which can be described as

$$GI(t) = \sum_{i=1}^{M} P(t|C_i)^2 P(C_i|t)^2 \tag{2}$$

where $P(t|C_i)$ is the probability of term t given presence of class C_i, $P(C_i|t)$is the probability of class C_i given presence of term t, respectively.

2.3. Distinguishing feature selector (DFS)

DFS is one of the recent successful feature selection methods for text classification and is also a global feature selection metric (Uysal & Gunal, 2012). The idea behind DFS is to select distinctive features while eliminating uninformative ones considering some pre-determined criteria. DFS can be expresses with the following formula:

$$DFS(t) = \sum_{i=1}^{M} \frac{P(C_i|t)}{P(\bar{t}|C_i) + P(t|\bar{C}_i) + 1} \tag{3}$$

where M is the number of classes, $P(C_i|t)$ is the conditional probability of class C_i given presence of term t, $P(\bar{t}|C_i)$ is the conditional probability of absence of term t given class C_i, and $P(t|\bar{C}_i)$ is the conditional probability of term t given all the classes except C_i.

2.4. Odds ratio (OR)

OR metric measures the membership and non-membership to a specific class with its nominator and denominator, respectively. The membership and non-membership scores are normalized by dividing them with each other (Mengle & Goharian, 2009). So, the nominator must be maximized and the denominator must be minimized in order to get a highest score according to the formula. As it can also be understood from the formula, the method is a one-sided metric because the logarithm function produces negative scores while the value of the fraction is between 0 and 1. In this case, the features having negative values point to negative features. The formula of OR can be expressed as

$$OR(t, C_i) = \log \frac{P(t|C_i)[1 - P(t|\bar{C}_i)]}{[1 - P(t|C_i)]P(t|\bar{C}_i)} \tag{4}$$

where M is the number of classes, $P(t|C_i)$ is the probability of term t given presence of class C_i, and $P(t|\bar{C}_i)$ is the conditional probability of term t given all the classes except C_i. In this paper, a simple smoothing method was applied in order to avoid division by zero errors and prevent the nominator become zero. These situations are valuable as they represent maximum membership and non-membership. So, a small value such as 0.01 was added to both nominator and denominator of the fraction.

3. IGFSS

In a classical global feature selection scheme for text classification, feature selection scores indicating the discriminatory powers of all terms in a given collection are calculated initially. Then, these terms are sorted according to their feature selection scores in descending order. After this sorting process, top-N features are included in the feature set as a final step (Guyon & Elisseeff, 2003) where N is usually an empirically determined number. IGFSS method aims to improve the performance of classification by modifying the abovementioned global feature selection process. For this purpose, a one-sided local feature selection method is integrated to the feature selection process besides an existing global feature selection method. So, IGFSS can be regarded as an ensemble method where the power of a global feature selection method and a one-sided local feature selection method are combined in a different manner. The flow of IGFSS method is given as below:

Stage 1. (Feature labeling)

- Calculate one-sided local feature selection scores of features for each class.
- Create a label set l for features including $m * 2$ class labels where m is the number of classes. While the first m class labels represent membership, the second m labels represent non-membership to these classes.
- For each feature, determine the highest local feature selection score regarding their absolute values and assign the associated class label from the label set l to the feature.

Stage 2. (Common global feature selection process)

- Calculate feature selection scores for features using one of the global feature selection metrics.
- Sort the features in descending order according to the scores and the sorted list is named as sl.

Stage 3. (Construction of the new feature set)

- Suppose that the size of the final feature set was given as fs and a set of negative feature ratios was determined as $nfrs$. The values in $nfrs$ may change from 0 to 1 with a specified pre-determined interval such as 0.1.

Table 1
A sample collection.

Document name	Content	Class
Doc 1	mouse cat wolf	C1
Doc 2	mouse cat dog horse	C2
Doc 3	mouse cat dog mice horse	C2
Doc 4	bat cow duck horse pelican	C3
Doc 5	bat cow horse pelican	C3
Doc 6	bat cow rat horse mice	C3

Table 2
Feature selection scores and membership degrees.

Feature	GI scores	OR scores (C1, C2, C3)	OR label	Positive/ Negative
Bat	1	−4.1109, −4.3307, 4.6151	C3	Positive
Cow	1	−4.1109, −4.3307, 4.6151	C3	Positive
Dog	1	−3.7136, 4.6151, −4.2146	C2	Positive
Wolf	1	4.6151, −3.2581, −3.5361	C1	Positive
Cat	0.5556	4.1109, 4.3307, −4.6151	C3	Negative
Mouse	0.5556	4.1109, 4.3307, −4.6151	C3	Negative
Horse	0.5200	−4.6151, 3.2581, 3.5361	C1	Negative
Pelican	0.4444	−3.7136, −3.9318, 3.8165	C2	Negative
Duck	0.1111	−3.0445, −3.2581, 2.4941	C2	Negative
Rat	0.1111	−3.0445, −3.2581, 2.4941	C2	Negative
Mice	0.0903	−3.7136, 0, −1.2929	C1	Negative

Table 3
Final feature sets obtained with two different methods.

Method	Final feature set	Distributions of class labels
GI	bat, cow, dog, wolf, cat, mouse	C1 (1), C2 (1), C3 (4)
GI + IGFSS	bat, dog, wolf, cat, horse, pelican	C1 (2), C2 (2), C3 (2)

- Iterate over the sorted list *sl* obtained in the previous stage and put the appropriate features in the final feature set *ffs*. Make the *ffs* equally representative for each class by using the feature labels determined in stage 1. At the end of this stage, *ffs* must contain equal number of features for each class considering a specific negative membership ratio value *nfr* inside *nfrs*.

Stage 4. (Conditional part)

- If the number of features in *ffs* is less than *fs*, finalize the feature selection process by adding missing amount of disregarded features having highest global feature selection scores to *ffs*.

As can be understood from the flow of IGFSS, in the worst case, all features needs to be traversed once and some of them may be traversed two times while constructing candidate feature set. Apart from the explanations above, a sample collection is provided in Table 1 in order to illustrate how IGFSS works. As seen from Table 1, there exist six documents consisting of 11 distinct terms in the sample collection. The features that are sorted according to their GI values and their corresponding OR scores are presented in Table 2. In this table, feature labels and their associated membership degrees are also given. Then, feature sets obtained by GI and GI based IGFSS are shown in Table 3 where the size of the feature set was determined as 6 and the *nfr* was set to 0.5. It is necessary to emphasize that the value of *nfr* is given as 0.5 in order to explain the flow of the IGFSS method better. So, it is not an empirically determined value.

In this sample scenario, there are two main points drawing attention about global feature selection methods. The first one is that the classes may not be represented almost equally in the final feature set. According to the sample scenario, while 6 features having higher GI scores are selected, each class is represented with 1, 2, and 4 features, respectively. The second point is that most of the feature selection methods do not concern with negative features too much. The term ‘wolf’ representing membership to class C1 was added to the feature set. On the other hand, the term ‘horse’ which is a good indicator of non-membership to class C1 was not included in the feature set. However, if we analyze the discriminative power of the term ‘wolf’ and ‘horse’ manually, it can easily be seen that they have similar discriminative powers about C1. According to GI score, the term ‘wolf’ is nearly two times important than ‘horse’. As so some studies in the literature refers, negative features are also valuable and a portion of them must be included in the final feature set. For this purpose, a negative feature ratio can be empirically determined. With the help of IGFSS, both classes are represented almost equally and the negative features such as the term ‘horse’ can be added to the final feature set. The sample collection and the final feature sets are provided to show how IGFSS method works. Actual performance of IGFSS on various benchmark datasets with distinct characteristics is thoroughly assessed in the experimental work.

4. Classification algorithms

In order to prove the efficacy of the proposed method, it was necessary to employ the classifiers commonly used for text classification research in the literature and proven to be significantly successful. For this purpose, linear support vector machine (Joachims, 1998) and naïve Bayes (Chen, Huang, Tian, & Qu, 2009) classifiers were utilized. A brief explanation about these methods is given in the next subsections.

4.1. Support vector machine (SVM)

SVM is one of the most effective classification algorithms in the literature and it has both linear and nonlinear versions. In this study, linear version of SVM, which is known as one of the mostly successful one especially for text classification, is employed (Uysal & Gunal, 2012). SVM looks for a decision surface that is maximally far away from any data point. The distance from the decision surface to the closest data point determines the margin of the classifier. The essential point of SVM classifier is the margin maximization concept (Joachims, 1998; Theodoridis & Koutroumbas, 2008). For this purpose, support vectors, which are the data points that lie at the border between the two classes, are detected. In order to handle multi-class classification problems, one of the two common approaches, namely one-against-all and one-against-one, can be adapted to convert two-class classification to multi-class case (Uysal & Gunal, 2012). In the experiments, LIBSVM classification toolbox (Chang & Lin, 2011) is used with the default parameter settings.

4.2. Naïve bayes (NB)

Naïve Bayes classifiers are a kind of simple probabilistic classifiers based on Bayes theorem which regards the features as independent from each other. Essentially, due to this independence assumption, a probability score is calculated by multiplying the conditional probabilities with each other in naïve Bayes classification. Although there are some widely-known event models such as Gaussian for Naïve Bayes classifiers, multinomial and multi-variate Bernoulli event models are widely accepted ones for text classification (Jiang, Cai, Zhang, & Wang, 2013). The flow of the naïve Bayes algorithm can be described as follows. Let document *d* consisting of a number of words is represented as (5) and its corresponding class label is assigned according to Bayes rule as (6).

$$d = w_1, w_2, w_3, \ldots, w_n \tag{5}$$

$$label(d) = \underset{c}{argMax}\ (P(Y = c) \prod_{i=1}^{n} P(w_i|Y = c)) \tag{6}$$

In this case, $P(Y = c)$ is the probability of class c and $P(w_i|Y = c)$is the probability of word w_i for a given class c. Multinomial and multi-variate Bernoulli event models differs in calculation of $P(w_i|Y = c)$ in (6). This probability is calculated as (7) and (8) according to multinomial and multi-variate Bernoulli event models, respectively.

$$P(w_i|Y = c) = \frac{tf_{w_i,c}}{|c|} \tag{7}$$

$$P(w_i|Y = c) = \frac{df_{w_i,c}}{N_c} \tag{8}$$

In the formulas, $tf_{w_i,c}$ is term frequency of w_i in class c, $|c|$ is the sum of term frequencies in class c, $df_{w_i,c}$ is document frequency of w_i in class c, and N_c is the total number of documents in class c. If the word w_i does not exist in the document d, the probability formula changes as (9) for the word $w_{i.}$

$$P(w_i|Y = c) = 1 - P(w_i|Y = c) \tag{9}$$

In this study, multi-variate Bernoulli event model is utilized for naïve Bayes classification.

5. Experimental work

In this section, an in-depth investigation was carried out to measure the performance of IGFSS against the individual performance of the three global feature selection methods. While one-sided local feature selection method utilized in the flow of IGFSS was OR, global feature selection methods employed in the experiments were IG, GI, and DFS. It should also be noted that stop-word removal and stemming (Porter, 1980) were used as the two pre-processing steps besides weighting terms with term frequency-inverse document frequency (TF-IDF). In order to validate the performance of IGFSS, three different datasets with varying characteristics and two different success measures were utilized to observe effectiveness of IGFSS method under different circumstances. In the following subsections, the utilized datasets and success measures are briefly described. Then, the characteristics of feature sets produced by the global feature selection methods are analyzed in order to show that the classes are not equally represented in the feature set. Finally, the experimental results are presented.

5.1. Datasets

In this study, three distinct datasets with varying characteristics were used for the assessment. The first dataset consists of the top-10 classes of the celebrated Reuters-21578 ModApte split (Asuncion & Newman, 2007). The second dataset is another popular benchmark collection namely WebKB (Craven, McCallum, PiPasquo, Mitchell, & Freitag, 1998) which is consisted of four classes. The third dataset is Classic3 whose class distribution is nearly homogenous among three classes (Uguz, 2011). All of these three datasets are widely used benchmark collections for text classification. The detailed information regarding those datasets is provided in Tables 4–6. It is obvious from these tables that Reuters dataset is highly imbalanced, that is, numbers of documents in each class are quite different. On the contrary, WebKB and Classic3 datasets are more balanced ones with closer number of documents per class. While Reuters dataset has its own training and testing split, WebKB and Classic3 datasets were manually divided into training and testing splits. For this purpose, 70% and 30% of documents were used as training and testing, respectively.

5.2. Success measures

The two success measures employed in this study are well known Micro-F1 and Macro-F1 (Manning et al., 2008; Uysal & Gunal, 2012).

Table 4
Reuters dataset.

Nos.	Class label	Training samples	Testing samples
1	Earn	2877	1087
2	Acq	1650	719
3	Money-fx	538	179
4	Grain	433	149
5	Crude	389	189
6	Trade	369	117
7	Interest	347	131
8	Ship	197	89
9	Wheat	212	71
10	Corn	181	56

Table 5
WebKB dataset.

No	Class label	Training samples	Testing samples
1	Course	651	279
2	Faculty	786	338
3	Project	352	152
4	Student	1148	493

Table 6
Classic3 dataset.

No	Class label	Training samples	Testing samples
1	Cisi	1021	439
2	Cran	978	420
3	Med	723	310

In micro-averaging, all classification decisions in the dataset are entirely considered without class discrimination. If the classes in a collection are biased, large classes would dominate small ones. Computation of Micro-F1 score can be formulated as

$$Micro - F1 = \frac{2 \times p \times r}{p + r} \tag{10}$$

where pair of (p, r) corresponds to precision and recall values, respectively, over all the classification decisions within the entire dataset not individual classes. However, in macro-averaging, F-measure is computed for each class within the dataset and then the average over all classes is obtained. In this way, equal weight is assigned to each class without regarding the class distributions. Computation of Macro-F1 can be formulated as

$$Macro - F1 = \frac{\sum_{k=1}^{C} F_k}{C}, \qquad F_k = \frac{2 \times p_k \times r_k}{p_k + r_k} \tag{11}$$

where pair of (p_k, r_k) corresponds to precision and recall values of class k, respectively.

5.3. Analysis of the feature sets produced by global feature selection methods

As pointed out in Section 3, the feature sets constructed by global feature selection methods may not represent all classes almost equally. In this part, the distributions of features to classes are analyzed for benchmark datasets. The profiles of feature sets obtained from Reuters-21578, WebKB, and Classic3 datasets by IG, GI, and DFS methods are investigated in the following figures. The labels that OR assigns to the features are used for this analysis. The amount of positive and negative features which shows the membership and non-membership to classes are also presented in the figures. This analysis is realized only on 250 features due to the fact that the feature size is the minimum dimension in the experiments. Fig. 1–3 show the class distributions of features obtained from Reuters dataset. It is clear that features and their corresponding non-membership distributions vary among classes for all three methods.

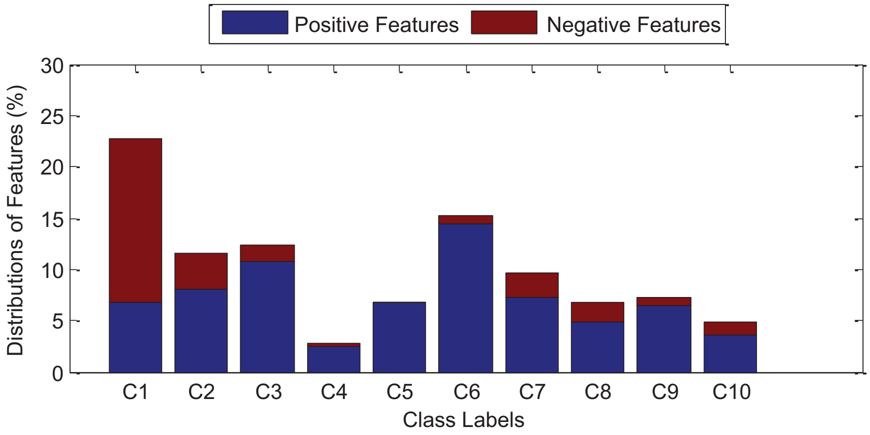

Fig. 1. Reuters: class distributions of features selected by IG.

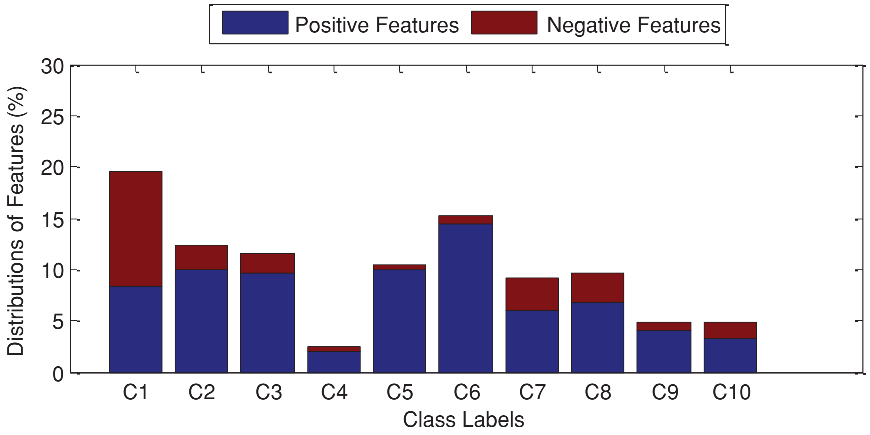

Fig. 2. Reuters: class distributions of features selected by GI.

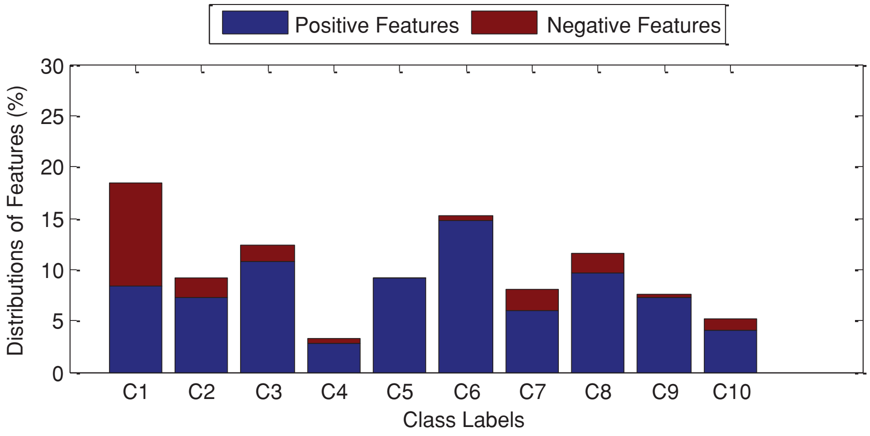

Fig. 3. Reuters: class distributions of features selected by DFS.

Fig. 4–6 show the class distributions of features selected by IG, GI, and DFS methods on WebKB dataset. As in the previous figures, distribution of the features and their corresponding non-memberships vary among classes for IG, GI, and DFS methods. It can be noted that the change is not proportional with class probabilities. While the class having maximum amount of documents is C4, C1 is the mostly presented class for WebKB dataset.

Fig. 7–9 show the class distributions of features selected by IG, GI, and DFS on Classic3 dataset. In this case, distributions of features to classes are more balanced but non-membership ratios change for different global feature selection methods. According to the figures, C2 is the mostly presented class in spite of not having the most training samples as it is valid for WebKB dataset.

5.4. Accuracy analysis

In this section, the individual performances of global feature selection methods and the proposed IGFSS method were compared. This comparison was carried out according to the maximum Micro-F1 and Macro-F1 values that these methods achieved. For IGFSS, the best performing negative feature ratio *nfr* and its corresponding Micro-F1 and Macro-F1 scores are presented. Bold cells in the tables indicate

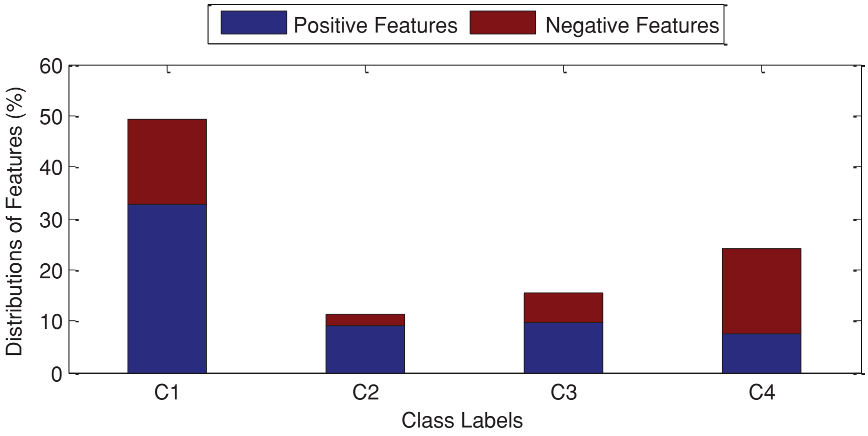

Fig. 4. WebKB: class distributions of features selected by IG.

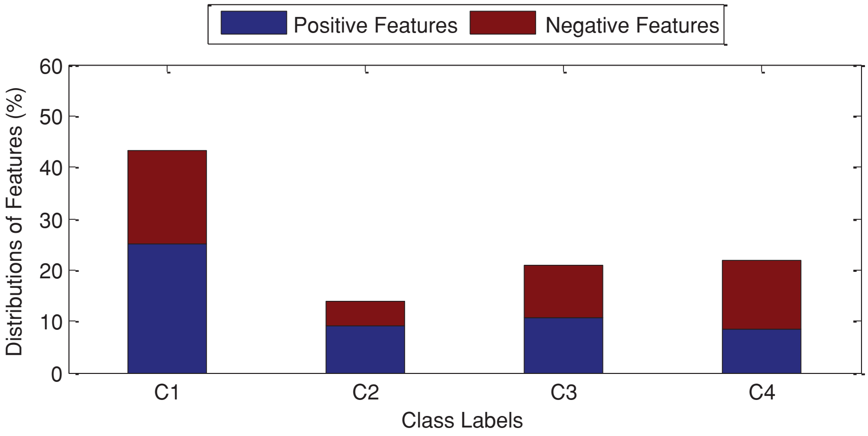

Fig. 5. WebKB: class distributions of features selected by GI.

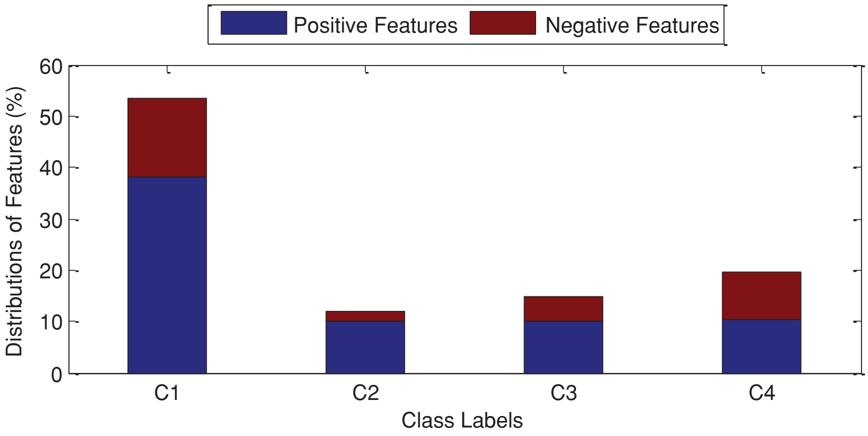

Fig. 6. WebKB: class distributions of features selected by DFS.

the maximum score for a specific method. Varying numbers of the features, which are selected by each selection method, were fed into SVM and NB classifiers. Table 7–9 show the Micro-F1 scores that were obtained on three different datasets with these two classifiers.

According to Tables 7–9, IGFSS method surpasses the individual performances of three different global feature selection methods in terms of Micro-F1. However, the value of the negative feature ratio *nfr* usually changes for different settings. The improvement on WebKB dataset seems more impressive than the other datasets. However, improvement on Classic3 dataset is lower than the others due to probably its structure. As pointed out in the previous subsection, it is a more balanced dataset than the others and this may have caused negative features to be ineffective. Low *nfr* values obtained for Classic3 dataset supports this idea. It is possible to say that NB classifier has improved better than SVM classifier. Besides, Tables 10–12 show the Macro-F1 scores that were obtained on three different datasets with these two classifiers.

According to Tables 10–12, IGFSS method outperforms the individual performances of three different global feature selection methods for all cases in terms of Macro-F1. It is also necessary to note that Micro-F1 and Macro-F1 values on Reuters dataset differs more than WebKB and Classic3 datasets because of being highly imbalanced.

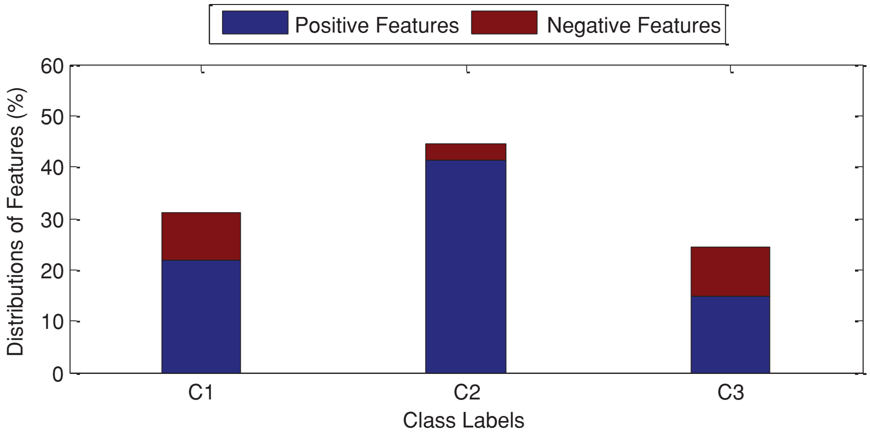

Fig. 7. Classic3: class distributions of features selected by IG.

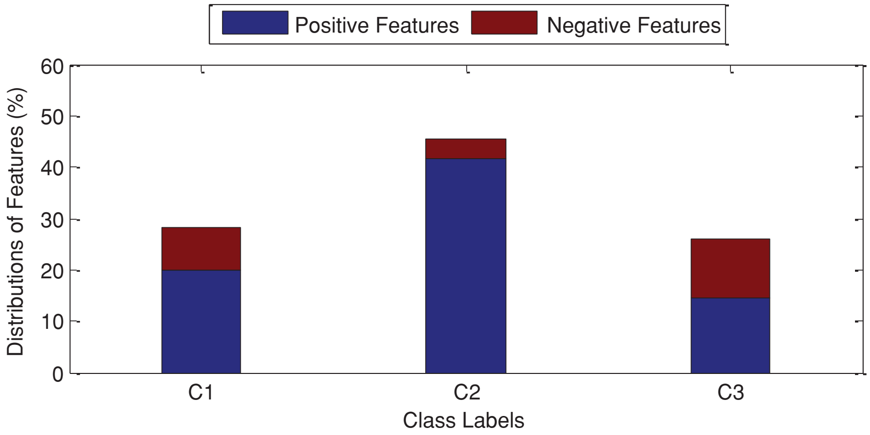

Fig. 8. Classic3: class distributions of features selected by GI.

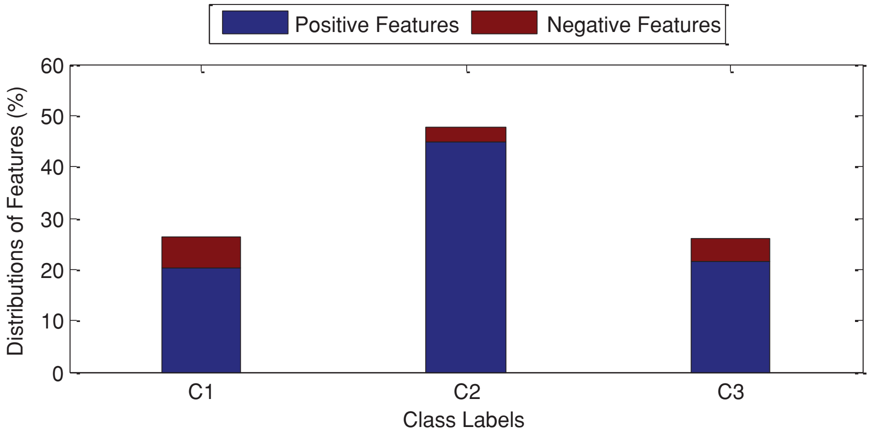

Fig. 9. Classic3: class distributions of features selected by DFS.

The values of negative feature ratio *nfr* obtaining higher scores are similar to the ones in the Micro-F1 case. The improvement on Reuters and WebKB datasets seems more outstanding than Classic3 dataset for Macro-F1 scores.

6. Conclusions

The main contribution of this study to the literature is to introduce an improved global feature selection scheme (IGFSS) for text classification. IGFSS is a generic solution for all of the filter-based global feature selection methods unlike most of the other approaches in the literature. As pointed out before, most of the studies in the literature are focused on providing some improvements on specific feature selection methods rather than providing a new generic scheme. IGFSS is an ensemble method which combines the power of a filter-based global feature selection method and a one-sided local feature selection method. The idea behind IGFSS is to make the feature set represent each class in the dataset almost equally. For this purpose, efficient feature ranking skills of global feature selection methods were combined with class membership and non-membership

Table 7
Micro-F1 scores (%) for Reuters dataset using (a) SVM (b) NB.

(a) Method		Micro-F1					
	nfr	250	300	350	400	450	500
IG	–	85.755	**86.006**	86.006	85.863	86.006	85.827
IG+IGFSS	0.6	85.361	**86.473**	86.150	86.294	86.114	86.006
GI	–	85.935	85.971	86.006	86.401	86.078	**86.437**
GI+IGFSS	0.3	85.648	85.791	86.329	86.437	**86.760**	85.935
DFS	–	85.899	85.899	**85.971**	85.791	85.899	85.791
DFS+IGFSS	0.8	85.002	86.258	**86.473**	86.258	86.114	85.863
(b) Method		Micro-F1					
IG	–	**83.531**	82.382	82.382	82.562	81.916	81.737
IG+IGFSS	0.3	84.105	84.284	84.320	84.212	**84.535**	84.033
GI	–	**84.535**	84.212	83.961	84.141	83.674	83.423
GI+IGFSS	0.3	85.109	**85.468**	84.822	84.966	84.356	84.571
DFS	–	**84.930**	84.284	84.033	83.889	83.602	83.100
DFS+IGFSS	0.4	84.607	85.181	**85.289**	84.679	84.787	84.751

Table 8
Micro-F1 scores (%) for WebKB dataset using (a) SVM (b) NB.

(a) Method		Micro-F1					
	nfr	250	300	350	400	450	500
IG	–	**82.013**	81.696	82.013	80.903	81.616	81.062
IG+IGFSS	0.7	83.597	**83.914**	81.933	81.854	81.696	79.794
GI	–	81.220	81.537	81.300	**83.439**	82.567	82.964
GI+IGFSS	0.7	**84.311**	83.043	82.013	82.567	82.726	81.696
DFS	–	**83.756**	83.677	82.409	81.379	80.586	79.952
DFS+IGFSS	0.7	**84.548**	82.726	82.250	81.696	81.062	80.586
(b) Method		Micro-F1					
IG	–	81.062	**81.220**	80.983	80.349	79.952	79.239
IG+IGFSS	0.2	83.122	**83.518**	83.043	82.647	81.458	80.983
GI	–	57.765	61.252	64.897	69.017	70.919	**72.583**
GI+IGFSS	0	**78.130**	77.655	77.734	77.338	76.941	76.624
DFS	–	**82.647**	81.616	82.250	81.854	80.745	80.666
DFS+IGFSS	0.3	**84.707**	83.360	82.964	82.567	83.043	82.567

Table 9
Micro-F1 scores (%) for Classic3 dataset using (a) SVM (b) NB.

(a) Method		Micro-F1					
	nfr	250	300	350	400	450	500
IG	–	95.723	96.407	96.920	97.348	96.920	**97.519**
IG+IGFSS	0.1	96.065	96.151	96.493	96.578	**97.605**	97.177
GI	–	94.440	95.552	96.920	96.322	**97.006**	96.920
GI+IGFSS	0.1	97.092	96.151	96.749	**97.177**	97.092	97.177
DFS	–	96.065	96.578	97.177	97.434	97.776	**97.947**
DFS+IGFSS	0.3	95.552	96.236	96.835	97.006	**98.033**	97.006
(b) Method		Micro-F1					
IG	–	97.263	97.605	98.289	98.375	98.460	**98.546**
IG+IGFSS	0	97.776	97.861	98.375	**98.973**	98.888	98.888
GI	–	96.835	97.947	**98.204**	97.947	98.118	98.118
GI+IGFSS	0	97.605	97.776	98.204	98.802	**98.888**	99.059
DFS	–	97.605	98.204	98.204	98.546	**98.802**	98.717
DFS+IGFSS	0	98.118	98.375	98.802	98.802	98.802	**98.973**

detection capability of one-sided local feature selection methods in a different manner. A specific negative feature ratio was determined while obtaining the new feature sets with these two methods. Using well-known benchmark datasets, classification algorithms and success measures, effectiveness of IGFSS was investigated and compared against the individual performance of filter-based global feature selection methods. The results of a thorough experimental analysis clearly indicate that IGFSS improved the performance of classification in terms of Micro-F1 and Macro-F1.

Despite its significant contribution, this proposed scheme has some limitations. In this study, IGFSS is applied to global feature selection methods for text classification. However, local feature

Table 10
Macro-F1 scores (%) for Reuters dataset using (a) SVM (b) NB.

(a) Method		Macro-F1					
	nfr	250	300	350	400	450	500
IG	–	65.384	65.040	**65.737**	65.184	64.715	64.798
IG+IGFSS	0.6	66.102	**67.533**	66.192	66.111	65.708	65.351
GI	–	64.761	65.358	**66.769**	66.062	65.410	65.958
GI+IGFSS	0.7	66.347	**67.277**	66.814	66.459	66.932	65.948
DFS	–	65.568	65.979	**66.170**	65.010	65.089	65.024
DFS+IGFSS	0.9	64.935	66.475	65.838	66.487	**67.076**	65.532

(b) Method		Macro-F1					
IG	–	**65.516**	64.389	64.625	64.293	64.044	63.753
IG+IGFSS	0.3	66.031	66.008	66.160	66.019	**66.641**	66.108
GI	–	65.783	**65.942**	65.044	65.302	65.155	65.126
GI+IGFSS	0.1	**68.390**	67.648	67.615	67.615	67.524	67.736
DFS	–	**66.770**	66.633	66.457	66.016	65.471	64.971
DFS+IGFSS	0.1	**68.514**	67.229	67.467	67.483	67.338	66.931

Table 11
Macro-F1 scores (%) for WebKB dataset using (a) SVM (b) NB.

(a) Method		Macro-F1					
	nfr	250	300	350	400	450	500
IG	–	80.715	80.215	**80.788**	79.998	80.545	79.870
IG+IGFSS	0.7	82.223	**82.785**	80.904	80.810	81.094	79.791
GI	–	79.760	80.663	80.314	**82.810**	81.630	82.189
GI+IGFSS	0.3	81.454	80.659	81.234	82.359	82.128	**83.413**
DFS	–	**82.559**	82.497	81.462	80.766	79.936	79.206
DFS+IGFSS	0.7	**83.268**	81.526	81.059	80.598	79.916	80.130

(b) Method		Macro-F1					
IG	–	81.164	**81.211**	80.890	80.172	79.962	79.244
IG+IGFSS	0.5	**83.707**	82.586	82.551	82.040	81.768	82.227
GI	–	59.787	63.112	66.285	69.773	71.271	**72.648**
GI+IGFSS	0	**77.220**	76.493	76.400	76.255	75.640	75.184
DFS	–	**83.152**	81.809	82.525	82.172	80.922	80.955
DFS+IGFSS	0.3	**84.782**	83.543	83.298	82.878	83.466	82.991

Table 12
Macro-F1 scores (%) for Classic3 dataset using (a) SVM (b) NB.

(a) Method		Macro-F1					
	nfr	250	300	350	400	450	500
IG	–	95.585	96.303	96.852	97.234	96.800	**97.423**
IG+IGFSS	0.1	95.999	96.115	96.392	96.490	**97.525**	97.074
GI	–	94.250	95.525	96.916	96.274	**96.952**	96.843
GI+IGFSS	0.2	95.902	96.008	96.612	95.670	96.477	**97.173**
DFS	–	96.006	96.532	97.137	97.371	97.724	**97.921**
DFS+IGFSS	0.3	95.488	96.230	96.768	96.969	**98.055**	97.033

(b) Method		Macro-F1					
IG	–	97.108	97.472	98.229	98.290	98.396	**98.488**
IG+IGFSS	0	97.662	97.749	98.304	**98.935**	98.860	98.857
GI	–	96.687	97.829	**98.107**	97.830	98.012	98.010
GI+IGFSS	0	97.489	97.674	98.119	98.751	98.844	**99.028**
DFS	–	97.511	98.139	98.120	98.489	**98.766**	98.674
DFS+IGFSS	0	98.046	98.320	98.766	98.754	98.754	**98.937**

selection metrics can be adapted to this scheme with addition of a globalization step in order to produce a unique score for features. Besides, in the experiments, odds ratio is utilized as one-sided feature selection method to extract negative features. As pointed out in the previous sections, odds ratio is known to produce excessive number of negative features. So, it is possible to employ high ratios for negative features in case the feature dimension is not a very high number.

Based on the limitation of this paper and the computational results, some potential directions for future research might be proposed. As an example, heuristic approaches may be integrated to IGFSS in order to detect a more appropriate ratio for negative features. Correspondingly, the impact of using varying negative feature ratio for classes may be examined. Apart from these, the integration of other global feature selection methods in the literature to IGFSS and ratio of probable performance improvements still remain as an interesting future work.

References

Asuncion, A., & Newman, D. J. (2007). *UCI Machine Learning Repository*. Irvine, CA: University of California, Department of Information and Computer Science.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*, 1–27.

Chen, J., Huang, H., Tian, S., & Qu, Y. (2009). Feature selection for text classification with Naïve Bayes. *Expert Systems with Applications, 36*, 5432–5435.

Craven, M., McCallum, A., PiPasquo, D., Mitchell, T., & Freitag, D. (1998). Learning to extract symbolic knowledge from the world wide web. In: DTIC Document.

Dara, J., Dowling, J. N., Travers, D., Cooper, G. F., & Chapman, W. W. (2008). Evaluation of preprocessing techniques for chief complaint classification. *Journal of Biomedical Informatics, 41*, 613–623.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research, 3*, 1289–1305.

Gunal, S. (2012). Hybrid feature selection for text classification. *Turkish Journal of Electrical Engineering & Computer Sciences, 20*, 1296–1311.

Gunal, S., Ergin, S., Gulmezoglu, M. B., & Gerek, O. N. (2006). On feature extraction for spam e-mail detection. *Lecture Notes in Computer Science, 4105*, 635–642.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182.

Idris, I., & Selamat, A. (2014). Improved email spam detection model with negative selection algorithm and particle swarm optimization. *Applied Soft Computing, 22*, 11–27.

Jiang, L., Cai, Z., Zhang, H., & Wang, D. (2013). Naive Bayes text classifiers: a locally weighted learning approach. *Journal of Experimental & Theoretical Artificial Intelligence, 25*, 273–286.

Joachims, T. (1997). A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In *Proceedings of the 14th international conference on machine learning* (pp. 143–151). Nashville, USA

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. In C. Nédellec, & C. Rouveirol (Eds.), *Proceedings of the 10th european conference on machine learning: vol. 1398* (pp. 137–142).

Lee, C., & Lee, G. G. (2006). Information gain and divergence-based feature selection for machine learning-based text categorization. *Information Processing & Management, 42*, 155–165.

Manning, C. D., Raghavan, P., & Schutze, H. (2008). *Introduction to information retrieval*. New York, USA: Cambridge University Press.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: a survey. *Ain Shams Engineering Journal, 5*, 1093–1113.

Mengle, S. S. R., & Goharian, N. (2009). Ambiguity measure feature-selection algorithm. *Journal of the American Society for Information Science and Technology, 60*, 1037–1050.

Ogura, H., Amano, H., & Kondo, M. (2010). Distinctive characteristics of a metric using deviations from Poisson for feature selection. *Expert Systems with Applications, 37*, 2273–2281.

Ogura, H., Amano, H., & Kondo, M. (2011). Comparison of metrics for feature selection in imbalanced text classification. *Expert Systems with Applications, 38*, 4978–4989.

Pietramala, A., Policicchio, V. L., & Rullo, P. (2012). Automatic filtering of valuable features for text categorization. In *Proceedings of the advanced data mining and applications* (pp. 284–295). Springer Berlin Heidelberg.

Pinheiro, R. H. W., Cavalcanti, G. D. C., Correa, R. F., & Ren, T. I. (2012). A global-ranking local feature selection method for text categorization. *Expert Systems with Applications, 39*, 12851–12857.

Pinheiro, R. H. W., Cavalcanti, G. D. C., & Ren, T. I. (2015). Data-driven global-ranking local feature selection methods for text categorization. *Expert Systems with Applications, 42*, 1941–1949.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14*, 130–137.

Rill, S., Reinel, D., Scheidt, J., & Zicari, R. V. (2014). PoliTwi: early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis. *Knowledge-Based Systems, 69*, 24–33.

Saraç, E., & Özel, S. A. (2014). An ant colony optimization based feature selection for web page classification. *The Scientific World Journal*, 1–16.

Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y., & Wang, Z. (2007). A novel feature selection algorithm for text categorization. *Expert Systems with Applications, 33*, 1–5.

Taşcı, Ş., & Güngör, T. (2013). Comparison of text feature selection policies and using an adaptive framework. *Expert Systems with Applications, 40*, 4871–4886.

Theodoridis, S., & Koutroumbas, K. (2008). *Pattern recognition* (4 ed.). Academic Press.

Uguz, H. (2011). A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. *Knowledge-Based Systems, 24*, 1024–1032.

Uysal, A. K., & Gunal, S. (2012). A novel probabilistic feature selection method for text classification. *Knowledge-Based Systems, 36*, 226–235.

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing & Management, 50*, 104–112.

Uysal, A. K., Gunal, S., Ergin, S., & Gunal, E. S. (2013). The impact of feature extraction and selection on SMS spam filtering. *Elektronika ir Elektrotechnika (Electronics and Electrical Engineering)*.

Vicient, C., Sánchez, D., & Moreno, A. (2013). An automatic approach for ontology-based feature extraction from heterogeneous textualresources. *Engineering Applications of Artificial Intelligence, 26*, 1092–1106.

Wang, Y., Liu, Y., Feng, L., & Zhu, X. (2015). Novel feature selection method based on harmony search for email classification. *Knowledge-Based Systems, 73*, 311–323.

Yang, B., Zhang, Y., & Li, X. (2011). Classifying text streams by keywords using classifier ensemble. *Data & Knowledge Engineering, 70*, 775–793.

Yang, J., Qu, Z., & Liu, Z. (2014). Improved feature-selection method considering the imbalance problem in text categorization. *The Scientific World Journal*, 1–17.

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning* (pp. 412–420). Nashville, TN, USA

Zhang, C., Wu, X., Niu, Z., & Ding, W. (2014). Authorship identification from unstructured texts. *Knowledge-Based Systems, 66*, 99–111.

Zheng, Z., & Srihari, R. (2003). Optimally combining positive and negative features for text categorization. In *Proceedings of the ICML, 3*, 1–8.

Zheng, Z., Wu, X., & Srihari, R. (2004). Feature selection for text categorization on imbalanced data. *ACM SIGKDD Explorations Newsletter, 6*, 80–89.